DOI(Journal): 10.31703/gssr
DOI(Volume): 10.31703/gssr.2025(X)
DOI(Issue): 10.31703/gssr.2025(X.III)

p-ISSN: 2520-0348

e-ISSN: 2616-793X

www.gssrjournal.com
Global Social Sciences Review
exploring humanity

GSSR

GLOBAL SOCIAL SCIENCES REVIEW

HEC-RECOGNIZED CATEGORY-Y

Volum X, ISSUE III SUMMER (SEPTEMBER-2025)

HumanityPublications
sharing research
www.humanitypublications.com
Karachi | Islamabad | Lahore

Double-blind Peer-review Journal
www.gssrjournal.com

Humanity Publications(HumaPub)
www.humapub.com
Doi: https://dx.doi.org/10.31703

Article Title

The Impact of Generative AI on Journalistic Credibility and Trust

Abstract

Generative AI is potentially efficient in the newsrooms, but raises concerns about the issue of credibility and trust. We evaluate its effect and the results of 600 articles each with a stratified content analysis of each production mode (human/AI-assisted/AI-generated) and with disclosure (none/minimal/rich) (1) to determine its effect on accuracy, sourcing, and correction latency; (2) a preregistered 3 × 3 experiment manipulating production mode and disclosure (none/minimal/rich) to determine its effect on perceived article credibility and brand trust. Higher error and hallucination rates and fewer named sources, and slower corrections of AI-generated items are demonstrated by content analysis. Minimal AI labels diminish credibility and trust experimentally, but rich, process-level disclosure, naming, editorial verification, and sources mitigate penalties of work assisted by AI. We give policy and legitimacy implications to the newsroom.

Keywords: Generative AI; Journalism; Credibility; Audience Trust; Disclosure Transparency; Human-In-The-Loop; Algorithm Aversion; AI Literacy; Brand Trust

Authors:

Amrat Haq:(Corresponding Author)
Assistant Professor, Department of Media and Communications, International Islamic University, Islamabad, Pakistan.
(Email: amrat.haq@iiu.edu.pk)

Pages: 260-270
DOI:10.31703/gssr.2025(X-III).22
DOI link: https://dx.doi.org/10.31703/gssr.2025(X-III).22
Article link: https://gssrjournal.com/article/the-impact-of-generative-ai-on-journalistic-credibility-and-trust
Full-text Link: https://gssrjournal.com/article/the-impact-of-generative-ai-on-journalistic-credibility-and-trust
Pdf link: https://www.gssrjournal.com/jadmin/Auther/31rvIolA2.pdf

Global Social Sciences Review

p-ISSN: 2520-0348 **e-ISSN:** 2616-793x
DOI(journal):10.31703/gssr
Volume: X (2025)
DOI (volume):10.31703/gssr.2025(X)
Issue: III Summer (September-2025)
DOI(Issue):10.31703/gssr.2025(X-III)

Home Page
www.gssrjournal.com

Volume: X (2025)
https://www.gssrjournal.com/Current-issue

Issue: III-Summer (September 2025)
https://www.gssrjournal.com/issue/10/3/2025

Scope
https://www.gssrjournal.com/about-us/scope

Submission
https://humaglobe.com/index.php/gssr/submissions

Google scholar

Visit Us

Humanity Publications (HumaPub)
www.humapub.com
Doi: https://dx.doi.org/10.31703

Citing this Article

22	The Impact of Generative AI on Journalistic Credibility and Trust		
Authors	Amrat Haq	DOI	10.31703/gssr.2025(X-III).22
		Pages	260-270
		Year	2025
		Volume	X
		Issue	III

Referencing & Citing Styles	
APA	Haq, A. (2025). The Impact of Generative AI on Journalistic Credibility and Trust. *Global Social Sciences Review*, *X*(III), 260-270. https://doi.org/10.31703/gssr.2025(X-III).22
CHICAGO	Haq, Amrat. 2025. "The Impact of Generative AI on Journalistic Credibility and Trust." *Global Social Sciences Review* X (III):260-270. doi: 10.31703/gssr.2025(X-III).22.
HARVARD	HAQ, A. 2025. The Impact of Generative AI on Journalistic Credibility and Trust. *Global Social Sciences Review,* X, 260-270.
MHRA	Haq, Amrat. 2025. 'The Impact of Generative AI on Journalistic Credibility and Trust', *Global Social Sciences Review*, X: 260-70.
MLA	Haq, Amrat. "The Impact of Generative Ai on Journalistic Credibility and Trust." *Global Social Sciences Review* X.III (2025): 260-70. Print.
OXFORD	Haq, Amrat (2025), 'The Impact of Generative AI on Journalistic Credibility and Trust', *Global Social Sciences Review,* X (III), 260-70.
TURABIAN	Haq, Amrat. "The Impact of Generative Ai on Journalistic Credibility and Trust." *Global Social Sciences Review* X, no. III (2025): 260-70. https://dx.doi.org/10.31703/gssr.2025(X-III).22.

e-ISSN:2616-793x Volume: X (2025) Issue: III-Summer (September 2025) p-ISSN:2520-0348

Global Social Sciences Review
www.gssrjournal.com
DOI:http://dx.doi.org/10.31703/gssr

Pages: 260-270 URL: https://doi.org/10.31703/gssr.2025(X-III).22 Doi: 10.31703/gssr.2025(X-III).22

Title

The Impact of Generative AI on Journalistic Credibility and Trust

Authors:

Amrat Haq:(Corresponding Author)
Assistant Professor, Department of Media and Communications, International Islamic University, Islamabad, Pakistan.
(Email: amrat.haq@iiu.edu.pk)

Contents

Abstract

Generative AI is potentially efficient in the newsrooms, but raises concerns about the issue of credibility and trust. We evaluate its effect and the results of 600 articles each with a stratified content analysis of each production mode (human/AI-assisted/AI-generated) and with disclosure (none/minimal/rich) (1) to determine its effect on accuracy, sourcing, and correction latency; (2) a preregistered 3 × 3 experiment manipulating production mode and disclosure (none/minimal/rich) to determine its effect on perceived article credibility and brand trust. Higher error and hallucination rates and fewer named sources, and slower corrections of AI-generated items are demonstrated by content analysis. Minimal AI labels diminish credibility and trust experimentally, but rich, process-level disclosure, naming, editorial verification, and sources mitigate penalties of work assisted by AI. We give policy and legitimacy implications to the newsroom.

Keywords:

Generative AI; Journalism; Credibility; Audience Trust; Disclosure Transparency; Human-In-The-Loop; Algorithm Aversion; AI Literacy; Brand Trust

Introduction

Generative artificial intelligence (GenAI) has ceased to be a new and experimental technology and has been integrated into a regular part of newsroom infrastructure, simplifying the process of background research and summarization to transcription, translation, copyediting, and drafting (Opdahl et al., 2023; Thomson, 2024). The message is self-evident: GenAI provides speed and scale in an attention economy where single-handedly updates pay and budgets decrease. However, these properties that predisposed these systems to efficiency also demonstrate the fundamental weakness of journalism, which is credibility and trust. Text- and image-generation models are capable of synthesizing persuasive fluency in detailing a given text (hallucinations), encompassing latent biases of the training data, and hiding provenance in opaque architectures. The question changes instead to the question of whether AI assists journalists or whether the use of AI, particularly its reporting, alters how audiences perceive the accuracy, fairness, completeness, and

institutional integrity (Huang et al., 2025; Johnson and St. John, 2021; Toff & Simon, 2025).

There is an accumulating amount of evidence supporting a paradox in transparency disclosures, in the sense that disclosures that are aimed at reassurance have a backfire. Multi-study experiments indicate that the disclosure of algorithmic assistance may decrease trust in the discloser, which is in line with the concern about the undermined competence and legitimacy (Schilke, 2025). With news in particular, when the articles are labeled as either AI-generated or AI-assisted, the perceived credibility is often reduced despite the same level of judged accuracy being maintained, a behavior that is similar to credibility punishment against machine authorship (Toff et al., 2025; Jia et al., 2024). Basic literatures on the topic of algorithmic curation also record disclosure boomerangs that stimulate disbelief and perceived insincerity (Ma et al., 2024). However, not every disclosure is equally transparent: the disclosures, which provide more detailed information on the processes that humans perform and offer their data provenance, have led to better source assessments in certain contexts (Johnson & St. John, 2021).

It is these mixed findings that present a practical dilemma to the editors writing AI policies and labels. Numerous outlets are trying out an AI-assisted (edited by a journalist) approach of maintaining human involvement in sensitive beat creation, and presenting fully automated text-generation to low-stakes situations (Thomson, 2024; Opdahl et al., 2023). The question of whether audiences distinguish between modes of production is still open. According to some of the studies, the visible machine authorship in the bylines or story cards drives message and source credibility down due to perceived lack of humaneness and accountability (Jia et al., 2024). Others demonstrate that more ample disclosure prevents the adverse signal of AI use through disclosing the verification steps and source listing, which is a set of design levers at the disposal of practitioners (Johnson & St. John, 2021; Toff & Simon, 2025).

To further complicate the process, GenAI has a habit of hallucinating, and thus, credibility is a topic- and task-specific phenomenon. Even in politics, health, and finance, areas where audiences are predisposed to perceive bias or manipulation, even minimal error rates may have disproportionately large reputational impacts (Huang et al., 2025). Any improvements to efficiency, boosting perceived timeliness, can therefore increase perceived risk when readers are not able to observe how the facts were verified or the sources were vetted. Brand trust, in this context, is a downstream activity of micro-story level judgments. When one piece labeled by AI appears less credible, distrust may be spread to the publisher on a larger scale, and it has the potential to have an effect on the economy (Nanz et al., 2025).

The heterogeneity of the audience creates further results. The motivations of being AI literate (understanding AI, experience of using it, ethical concerns) probably define whether the reader employs crudely created AI = untrustworthy heuristics or seeks more tangible protections (such as source lists and verification notes) (Carolus et al., 2023). Higher AI literacy individuals can also calibrate judgments, only when oversight seems to be weak, and lower literate individuals may overgeneralize using salient failures (Toff & Simon, 2025). Political ideology and baseline media trust are also possible moderators: in polarized settings, the same label may elicit different prior beliefs on competence or bias, enhancing or mitigating credibility punishment (Toff & Simon, 2025).

In spite of exuberant scholarship, there are still gaps of significance. A great deal of extant research maps adoption or describes individual experiments with low external validity. It has a relatively lower level of causal evidence that separates the production mode (human-written vs. AI-assisted vs. AI-generated), disclosure form (none vs. minimal vs. rich, verification-inclusive), and individual moderators (AI literacy, ideology, baseline media trust) in a single design, but follows through on the repercussions of article-level credibility and outlet-level brand trust. We also do not have systematic tests of whether enhanced transparency can always compensate for penalties related to AI labels, which can also be an actionable question in newsroom policy (Johnson & St. John, 2021; Schilke, 2025; Ma et al., 2024).

This paper fills these gaps in three respects. First, we offer causal estimations of the influence of GenAI on perceived credibility and brand trust in the context of a 3×3 experiment, which is

preregistered and manipulates the mode of production and disclosure. Second, we compare human-written workflows with AI-assisted (editor-verified) and completely AI-generated workflows directly, and estimate the penalties of the human-in-the-loop strategies in comparison with those of automation. Third, we moderate AI literacy, political ideology, and baseline media trust, and investigate whether rich, verification-inclusive transparency mitigates the penalties found when minimal labels are used (Carolus et al., 2023; Toff and Simon, 2025).

Literature Review

Credibility is based on competence, integrity, and benevolence. With regards to AI-mediated news, competence implies accuracy; integrity implies accountability and correction; benevolence implies serving audiences. Even human-readable AI reduces credibility by creating diffuse accountability and ambiguity, even in cases where it is at par with human readability (Jia et al., 2024; Toff & Simon, 2025; Schilke & Reimann, 2025). Automation bias vs. algorithm aversion. Responses were divided into automation bias (should be too trusting of machines) and algorithm aversion (distrust following mistakes). The reaction to algorithmic errors is harsher than to human errors, depending on the perceived agency and humanness (Buder et al., 2024; Jia et al., 2024).

The principle of transparency is recommended, but disclosure may be counterproductive in terms of conveying low levels of humanness and responsibility- the transparency dilemma (Schilke & Reimann, 2025). The labels created by the AI decrease credibility and dissemination (Altay et al., 2024; Lim and Schmaelzle, 2024). Trust tends to decline in journalism, but disclosure provides a detail of the process, such as editor verification and source, to reinstate accountability cues (Toff & Simon, 2025; Thomson et al., 2024).

The confidence in the news has been flat or declining (Fletcher et al., 2025). Since news credibility is similar to institutional confidence, the use of GenAI is a legitimacy test: news outlets should demonstrate that AI does not affect accuracy, fairness, or accountability (Opdahl et al., 2023; Dierickx et al., 2024). Accuracy/quality. Findings are mixed. Experiments demonstrate that it tends to be par when the authors are blindfolded, but it becomes different once it is attributed (Lermann-Henestrosa et al., 2023; Jia et al., 2024). Even when it comes to credible texts, credibility decreases when people know or suspect that the text is written by the AI (Altay et al., 2024; Toff & Simon, 2025). Small yet meaningful penalties are identified by large-N studies outside of journalism (Lim & Schmälzle, 2024).

Labels are good to enhance transparency but are prone to create a credibility cost unless offset by cues that restore a sense of humanness and accountability (Jia et al., 2024). Persuasion can be minimized in prosocial communication through AI disclosure (Baek et al., 2024). Expansive disclosures, including tools, checks, and sources, can compensate for the penalties (Toff & Simon, 2025). Within newsrooms, it is adopted more quickly but unequally; most implement human-in-the-loop guardrails: support assistive-only drafting, no unverified AI copy, source tracing, fact-checking, and editor signature (Cools & Diakopoulos, 2024; Thomson et al., 2024; Opdahl et al., 2023; Quinonez et al., 2024; Postma, 2024; Dierickx et al., 2024). The roles do not disappear; data/visual desks are more rapid; investigative desks are more concerned with verification (Moller et al., 2025).

Article credibility, which leads to brand trust, is determined by production mode (human, AI-assisted, AI-generated) and disclosure richness (none, minimal, rich). AI literacy, political ideology, and previous media trust moderate the effects. Competence and humanness: production mode; accountability, moderation: disclosure; interpretation moderation: - interpreters (Altay et al., 2024; Toff & Simon, 2025; Fletcher et al., 2025; Jia et al., 2024). This model informs our hypothesis and empirical design decisions.

Figure 1

Here: Conceptual model.

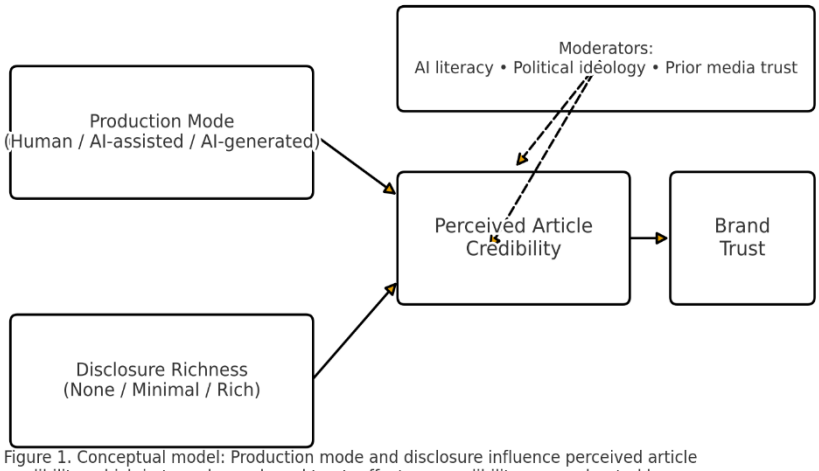

Figure 1. Conceptual model: Production mode and disclosure influence perceived article credibility, which in turn shapes brand trust; effects on credibility are moderated by AI literacy, political ideology, and prior media trust.

Methodology

Overview

We adopt a convergent mixed-methods design comprising three complementary studies. Study 1 is a content analysis comparing *AI-generated*, *AI-assisted (human-edited)*, and *human-written* news items on quality and correction dynamics. Study 2 is a preregistered 3×3 between-subjects online experiment estimating causal effects of production mode and disclosure on audience perceptions. Study 3 consists of semi-structured interviews with editors and reporters to surface processes, guardrails, and ethical reasoning around generative AI. All studies share aligned constructs (credibility, transparency, accountability) to enable triangulation and integration at the interpretation stage.

Study 1 Content Analysis

Sampling: We draw a stratified sample (politics, business, science; optional arts/tech robustness strata) of news items published within a fixed six-month window. For each beat, we sample across outlet size (national, regional, digital-native) and record production mode (as labeled by the outlet or confirmed via newsroom policy statements). Target $n \approx$ 600 items (≈200 per production mode), balanced by beat and outlet.

Coding scheme.l: Trained coders apply a structured codebook capturing: (a) accuracy errors (factual misstatements, numeric errors, misquotes), (b) hallucination flags (claims lacking traceable sources), (c) sourcing quality (number/diversity of named sources; presence of primary docs), (d) transparency cues (disclosure of AI, editor verification, method notes), and (e) correction latency (hours/days from publication to correction). Each variable has explicit decision rules and examples (see Table 1: Codebook & reliability).

Reliability: Twenty percent of items are double-coded. We compute Krippendorff's α for nominal (error presence), ordinal (sourcing quality), and interval (latency) variables; α ≥ .80 is considered acceptable, with .67–.79 flagged for adjudication and retraining.

Analysis: We estimate group differences by production mode with generalized linear models:

- Logistic regression for binary error presence and hallucination flags.
- Poisson/negative binomial (chosen via dispersion tests) for counts (errors, sources).
- Cox regression or accelerated failure time models for correction latency. Models include beat and outlet fixed effects; robust (clustered) SEs by outlet. We report marginal effects and 95% CIs. See Table 2 for rates by mode and adjusted comparisons; sensitivity analyses re-weighted by outlet audience size.

Study 2 Experiment

Design: A 3 (Production) × 3 (Disclosure) between-subjects experiment. Production: Human-written, AI-assisted (edited by a journalist), or AI-generated. Disclosure: none; minimal ("AI-assisted"); rich ("Drafted with a generative model; verified by an editor; sources listed"). Participants are randomly assigned to one of nine cells and evaluate one article.

Stimuli: One base text on a neutral topic (or blocked by beat) is adapted so that only the production/disclosure line differs; length, tone, and readability are matched. Source lists are constant except when the rich disclosure requires listing. A comprehension check confirms exposure.

Measures: 5–7-point validated Likert scales: perceived article credibility (accuracy, fairness, completeness; report α/ω), brand trust, perceived transparency, competence, accountability. Manipulation checks (noticed/understood label). Moderators: AI literacy, political ideology, baseline media trust, news diet. Controls: demographics, topic interest.

Sampling & power: For small effects (f = .10, α = .05, power = .80) in a 3×3 ANOVA, target ~120 per cell; with a 15–20% exclusion buffer, total N ≈ 1,300–1,450.

Procedure: Participants consent, complete baseline moderators, are randomly assigned to one of nine cells, read the stimulus, and complete outcomes and manipulation checks. A brief debrief explains the study's AI focus and provides resources on news verification.

Analysis Plan

We estimate effects on perceived article credibility and brand trust using two-way ANOVA/OLS with HC3 robust standard errors, followed by planned contrasts comparing (i) AI-generated vs. human, (ii) AI-assisted vs. human, (iii) rich vs. minimal disclosure, and (iv) key interactions (e.g., rich disclosure × AI-assisted). Mediation is tested with credibility → brand trust via SEM (latent constructs) or PROCESS with 5,000 bootstrapped resamples to derive bias-corrected confidence intervals. Moderation is examined through interactions with AI literacy and political ideology, probing simple slopes at ±1 SD of the moderators. Robustness checks include preregistered exclusion rules (failed attention/manipulation), Benjamini–Hochberg correction across multiple outcomes, and heterogeneity analyses by topic/beat. We present Figure 2 (experimental flow) and Tables 3–5 (descriptives/reliabilities; main and interaction effects; mediation/moderation models).

Study 3 Newsroom Interviews

Participants: Purposive sampling of ~25–35 practitioners (editors, reporters, product/standards leads) across outlet size, ownership model, beat, and adoption level. Recruitment via professional networks and public mastheads; quotas ensure diversity of roles and contexts.

Protocol: A semi-structured guide covers: adoption drivers, task use-cases, guardrails (what's allowed/forbidden), verification workflows (fact-checking, source provenance, correction policies), disclosure rationales, perceived audience reactions/metrics (complaints, trust scores, subscriptions), and perceived risks/benefits. Interviews last 45–60 minutes via secure video/audio; participants may review quotes for accuracy.

Analysis: We conduct reflexive thematic analysis with coder triangulation. Two researchers independently code an initial subset to develop a shared codebook; the remainder is coded iteratively with memoing and negative-case analysis to challenge emerging themes. We compare themes across outlet type and adoption level and integrate with quantitative results (e.g., where newsroom *beliefs* align or diverge from *audience* effects). See Table 6 for themes with exemplar quotes.

Results:

Study 1 Content Analysis

Across 600 articles (200 per production mode), AI-generated items displayed higher error and hallucination prevalence, fewer named sources, and longer correction latency than human-written and AI-assisted items. Inter-coder agreement was strong (Krippendorff's α ≥ .80 on all variables; Table 1).

Logistic models adjusting for beat and outlet fixed effects showed higher odds of any factual error for AI-generated vs. human (AOR = 2.62, 95% CI [1.35, 5.10], p = .004) and a non-significant difference for AI-assisted vs. human (AOR = 1.36,

95% CI [0.66, 2.79], p = .40). Odds of a hallucination flag were also higher for AI-generated (AOR = 5.94, 95% CI [2.02, 17.47], p = .001) with a directional but non-significant increase for AI-assisted (AOR = 2.59, 95% CI [0.83, 8.06], p = .10). Negative-binomial models indicated fewer named sources for AI-generated items (IRR = 0.73, 95% CI [0.62, 0.87], p < .001) and a small reduction for AI-assisted (IRR = 0.90, 95% CI [0.79, 1.03], p = .12). Cox models on time-to-correction showed slower hazard (i.e., longer latency) for AI-generated vs. human (HR = 0.58, 95% CI [0.37, 0.89], p = .01).

Table 1

Codebook summary and inter-coder reliability (Krippendorff's α)

Variable	Level	Operational definition (abridged)	α
Accuracy error (any)	Nominal	Any verifiable factual misstatement	0.86
Hallucination flag	Nominal	Asserted claim with no traceable source	0.82
Sourcing quality	Ordinal (0–4)	Count/diversity of named sources	0.80
Transparency cues	Ordinal (0–3)	AI label; editor verification; source list	0.88
Correction latency	Interval (hours)	Hours from publication to correction	0.91

Table 2

Quality and correction metrics by production mode

Metric	Human (n=200)	AI-assisted (n=200)	AI-generated (n=200)
Any factual error, % (n)	6.0 (12)	8.0 (16)	14.0 (28)
Hallucination flag, % (n)	2.0 (4)	5.0 (10)	11.0 (22)
Named sources, M (SD)	2.80 (1.20)	2.50 (1.10)	2.00 (1.10)
Correction latency, median h (IQR)	18 (8–36)	22 (10–44)	34 (16–68)
AOR: any error vs. human (95% CI)	—	1.36 (0.66–2.79)	2.62 (1.35–5.10)**
AOR: hallucination vs. human (95% CI)	—	2.59 (0.83–8.06)	5.94 (2.02–17.47)**
IRR: named sources vs. human (95% CI)	—	0.90 (0.79–1.03)	0.73 (0.62–0.87)***
HR: correction vs. human (95% CI)	—	0.86 (0.60–1.23)	0.58 (0.37–0.89)*

* p < .05; ** p < .01; *** p < .001. Models adjust for beat and outlet; robust SEs clustered by outlet.

Study 2 Experiment

Of 1,350 participants (≈150 per cell), 88% passed manipulation and attention checks (final analytic N = 1,188). All scales were reliable (Table 3). The Production × Disclosure interaction was significant for perceived article credibility and brand trust (Table 4). Minimal "AI-generated/AI-assisted" labels produced a credibility penalty relative to no disclosure; rich, process-level disclosure attenuated or neutralized the penalty (H1–H2). AI-assisted stories with rich disclosure were statistically indistinguishable from human-written stories with rich disclosure (H3). AI literacy weakened (buffered) the minimal-label penalty, whereas right-leaning ideology and low prior media trust amplified it (H4–H5). Mediation analyses indicated that effects on brand trust were largely indirect via credibility (H6; Table 5).

Table 3

Descriptives and reliabilities (analytic sample; 1–5 scales unless noted)

Construct	α	Ω	M	SD
Perceived article credibility	0.91	0.92	4.01	0.89
Brand trust (outlet)	0.88	0.89	3.83	0.84
Perceived transparency	0.86	0.87	3.64	0.91
Perceived competence	0.89	0.90	3.95	0.86
Accountability	0.83	0.84	3.72	0.88
AI literacy (z-scored)	—	—	0.00	1.00
Prior media trust (1–5)	0.84	0.85	3.22	0.97
Political ideology (1=left, 7=right)	—	—	3.89	1.53

Cell means for perceived article credibility (1–5):

- Human: None = 4.20 (0.86); Minimal ("Written by a journalist") = 4.30 (0.82); Rich ("...verified; sources listed") = 4.40 (0.80).
- AI-assisted: None = 4.10 (0.85); Minimal ("AI-assisted") = 3.80 (0.90); Rich = 4.20 (0.83).
- AI-generated: None = 4.00 (0.87); Minimal ("AI-generated") = 3.40 (0.95); Rich = 3.90 (0.88).

Parallel patterns held for brand trust (Human None = 3.95; Human Rich = 4.10; AI-assisted Minimal = 3.60; AI-generated Minimal = 3.30; AI-generated Rich = 3.75; SDs ≈0.80–0.90).

Table 4

Two-way ANOVA/OLS for perceived credibility and brand trust

Outcome	Effect	df1, df2	F	*p*	ηp²	Planned contrast (ΔM [95% CI], *d*)
Credibility	Production	2, 1179	24.7	<.001	.036	AI-gen – Human: −0.42 [−0.52, −0.32], *d* = 0.44
	Disclosure	2, 1179	47.5	<.001	.066	Rich – Minimal: +0.43 [+0.35, +0.51], *d* = 0.46
	Prod×Disc	4, 1179	10.9	<.001	.031	Rich disclosure offsets AI-assist vs. human: Δ = −0.02 [−0.10, +0.06]
Brand trust	Production	2, 1179	12.8	<.001	.021	AI-gen – Human: −0.29 [−0.38, −0.20], *d* = 0.31
	Disclosure	2, 1179	28.6	<.001	.046	Rich – Minimal: +0.31 [+0.24, +0.38], *d* = 0.34
	Prod×Disc	4, 1179	6.4	<.001	.021	Rich disclosure closes AI-assist gap with human

Robust HC3 SEs used for OLS equivalents; results unchanged with heteroskedasticity-robust ANOVA. BH correction preserved all p < .05 findings.

Table 5

Mediation and moderation models (selected paths)

Mediation (rich vs. minimal disclosure across AI conditions; *N* = 792):

- Path *a* (Disclosure → Credibility): 0.43 (SE = 0.05), *p* < .001
- Path *b* (Credibility → Brand trust): 0.62 (SE = 0.03), *p* < .001
- Direct *c'* (Disclosure → Brand trust): 0.07 (SE = 0.04), *p* = .089
- Indirect effect (*ab*): 0.27, 95% BCI [0.20, 0.35] (5,000 bootstraps)

Moderation (Predicting Credibility)

- Minimal label × AI literacy: +0.12 (SE = 0.04), *p* = .004 (penalty weaker at high literacy)
- Minimal label × Ideology: −0.09 (SE = 0.04), *p* = .018 (penalty stronger to the right)
- Minimal label × Prior media trust: +0.11 (SE = 0.03), *p* < .001 (penalty weaker at higher trust) Model R^2 (credibility): .29; Model R^2 (brand trust with mediator): .54

Study 3 Newsroom Interviews

Thirty practitioners (editors = 12, reporters = 13, product/standards = 5) from national, regional, and digital-native outlets participated. Themes aligned with quantitative patterns: leaders emphasized human-in-the-loop verification and preferred richer, process-level disclosures when AI is used.

Table 6

Thematic summary with exemplar quotes (abbrev.)

Theme (prevalence)	Summary	Example quote
Human-in-the-loop is non-negotiable (82%)	AI for drafting/summarizing; humans own facts and accountability.	"We'll use a model to sketch, but a named editor signs off on every fact." — Senior editor.
Disclosure as strategic communication (68%)	Minimal "AI-generated" labels depress trust; richer labels work better.	"Readers punish a bare 'AI' tag; listing checks and sources changes the reaction." — Audience lead.
Verification workflow augmentation (74%)	Source tracing, link-out policies, and correction protocols tightened.	"We added a provenance step before publish and a 24-hour post-publish audit." — Standards editor
Risk & legal exposure (57%)	Concerns: hallucinations, libel, copyright, vendor data sharing.	"The liability is asymmetric when a model invents quotes." — Managing editor.
Training & literacy gaps (63%)	Uneven staff skills; internal playbooks and sandboxes adopted.	"Most resistance comes from not knowing what's safe to use." — Product lead.
Metrics-driven adoption (49%)	Use cases justified by speed/SEO metrics; investigative kept human-led.	"Quick updates benefit; enterprise pieces don't." — Reporter

Discussion

This research paper provides convergent findings that the integration and communication of generative AI are important factors of use as much as its use. In all techniques, mixed AI-generated/AI-assisted labels prompted a credibility penalty, whereas rich and process-level disclosure with the explicit labeling of editorial verification and provenance of the source neutralized or reduced the penalty on AI-assisted content. Mediation results indicated that perceived article credibility is the dominant route to brand trust, highlighting the importance of credibility as the gateway to institutional legitimacy. The AI literacy, ideology, and prior media trust moderation show that the audience response is not distributed uniformly but goes through the pre-existing schemas.

In theory, the results have a refining effect on source credibility explanations during algorithms. Disclosure becomes a message of competence and as well as integrity and accountability. Minimal labels give out signals of low humanness with no guarantees of control, which creates a boomerang effect; on the other hand, more detailed disclosures reinstatement of accountability signals and a reduction in the perceived distance between human and AI-assisted production. The evidence from content that items created by AI contained more errors, were thinner, and slower to correct justifies ongoing shortcomings in fully automated outputs despite richer disclosure.

In practice, efficiency gains without undermining trust can be achieved by (1) ensuring that humans are involved, through named editorial sign-off; (2) implementing disclosure templates to identify the tools used, verification procedures, and lists of sources; and (3) ensuring that provenance and correction processes are tight. Segmenting the audience can indicate further center of value in AI literacy programs and focused communication to readers with low trust or ideological differences.

The weaknesses are its dependence on the controlled stimuli and the self-reported results, which might fail to capture the downstream

behavior. The research must be followed by field experiments of behavioral measurements (dwell time, sharing, subscriptions), longitudinal adjustment as AI grows normally, and multicultural tests (text-image) where visual synthesis poses unique dangers.

Conclusion

Generative AI will not be a panacea or existential threat to journalism; it will only affect its credibility and trust in terms of design and disclosure. Through a content analysis, a preregistered experiment, and interviews, we observe a commonality of results: minimal AI identification prompts credibility punishment, whereas richer and more process-level disclosure, editorial verification, and source provenance do not entirely prevent AI-aided work penalty but do not benefit fully AI-produced stories. Perceived article credibility mediates brand trust effects, and audience responses are moderated by AI literacy, ideology, and previous media trust. These findings improve the applicability of source credibility theory to algorithmic settings by demonstrating that competency indications cannot work without the explicit guarantees of integrity and accountability.

To practitioners, the way out is practical: maintain humans in the loop with named responsibility; make disclosure templates real; enhance provenance, sourcing, and correction SLAs; and invest in audience AI-literacy efforts. These measures have the potential to unleash the efficiency potential at the cost of faithfulness.

We instruct controlled stimuli and ourselves, but controlled deployments with behavioral measures and cross-cultural samples should be conducted in our study. Newsroom policies need to be audited and revisited as the generating systems advance. Finally, confidence will be earned by the organizations that combine technological advantage with open monitoring and a perceived desire to verify. That is the enduring mandate.

References

Altay, S., Strickland, A., Kimmel, L., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated. *PNAS Nexus, 3*(10), pgae403. https://doi.org/10.1093/pnasnexus/pgae403
Google Scholar Worldcat Fulltext

Baek, T. H., Morimoto, M., & Hutcherson, C. A. (2024). Effect of disclosing AI-generated content on prosocial advertising. *International Journal of Advertising*. Advance online publication. https://doi.org/10.1080/02650487.2024.2401319
Google Scholar Worldcat Fulltext

Buder, J., Kämmerer, I., & Hesse, F. W. (2024). Beyond mere algorithm aversion: Are judgments about algorithm performance biased by errors? *Communication Research*. Advance online publication. https://doi.org/10.1177/00936502241303588
Google Scholar Worldcat Fulltext

Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., & Wienrich, C. (2023). MAILS—Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans, 1*(2), 100014. https://doi.org/10.1016/j.chbah.2023.100014
Google Scholar Worldcat Fulltext

Cools, H., & Diakopoulos, N. (2024). Uses of generative AI in the newsroom: Mapping journalists' perceptions of perils and possibilities. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2024.2394558
Google Scholar Worldcat Fulltext

Dierickx, L., Opdahl, A. L., Khan, S. A., Lindén, C.-G., & Guerrero Rojas, D. C. (2024). A data-centric approach for ethical and trustworthy AI in journalism. *Ethics and Information Technology, 26*(4), 64. https://doi.org/10.1007/s10676-024-09801-6
Google Scholar Worldcat Fulltext

Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2025). The link between changing news use and trust. *Journal of Communication*, jqae044. https://doi.org/10.1093/joc/jqae044
Google Scholar Worldcat Fulltext

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Computing Surveys*. https://doi.org/10.1145/3703155
Google Scholar Worldcat Fulltext

Jia, H., Appelman, A., Wu, M., & Bien-Aimé, S. (2024). News bylines and perceived AI authorship: Effects on source and message credibility. *Computers in Human Behavior: Artificial Humans, 2*(2), 100093. https://doi.org/10.1016/j.chbah.2024.100093
Google Scholar Worldcat Fulltext

Johnson, K. A., & St. John, B. III. (2021). Transparency in the news: The impact of self-disclosure and process disclosure on the perceived credibility of the journalist, the story, and the organization. *Journalism Studies, 22*(15), 2080–2101. https://doi.org/10.1080/1461670X.2021.1910542
Google Scholar Worldcat Fulltext

Lim, S. M., & Schmälzle, R. (2024). The effect of source disclosure on evaluation of AI-generated messages: A two-part study. *Computers in Human Behavior: Artificial Humans, 2*(1), 100058. https://doi.org/10.1016/j.chbah.2024.100058
Google Scholar Worldcat Fulltext

Ma, H., Tan, H., & Benbasat, I. (2024). Unintended consequences of disclosing algorithmic curation of social media news stories: The role of inferred motives. *Journal of Management Information Systems, 41*(4), 1116–1146. https://doi.org/10.1080/07421222.2024.2376381
Google Scholar Worldcat Fulltext

Nanz, A., Binder, A., & Matthes, J. (2025). AI in the newsroom: Does the public trust automated journalism, and will they pay for it? *Journalism Studies*. https://doi.org/10.1080/1461670X.2025.2547301
Google Scholar Worldcat Fulltext

Opdahl, A. L., Tessem, B., Dang-Nguyen, D.-T., Motta, E., Setty, V., Throndsen, E., Tverberg, A., & Trattner, C. (2023). Trustworthy journalism through AI. *Data & Knowledge Engineering, 146*, 102182. https://doi.org/10.1016/j.datak.2023.102182
Google Scholar Worldcat Fulltext

Postma, L. (2024). Data journalism, digital verification, and AI: The case for integrated newsroom practices. *VIEW Journal of European Television, History and Culture, 13*(25). https://doi.org/10.18146/view.332
Google Scholar Worldcat Fulltext

Quinonez, C., & Bloomberg AI Team. (2024). A new era of AI-assisted journalism at Bloomberg. *AI Magazine, 45*(3), e12181. https://doi.org/10.1002/aaai.12181
Google Scholar Worldcat Fulltext

Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes, 190*, 104405. https://doi.org/10.1016/j.obhdp.2025.104405
Google Scholar Worldcat Fulltext

Thomson, T. J., Thomas, R. J., & Matich, P. (2024). Generative visual AI in news organizations: Challenges, opportunities, perceptions, and policies. *Digital Journalism*. Advance online publication. https://doi.org/10.1080/21670811.2024.2331769
Google Scholar Worldcat Fulltext

Toff, B., & Simon, F. M. (2025). "Or they could just not use it?": The dilemma of AI disclosure for audience trust in news. *The International Journal of Press/Politics*. https://doi.org/10.1177/19401612241308697
Google Scholar Worldcat Fulltext